

DECAY BOUNDS FOR FUNCTIONS OF BANDED NON-HERMITIAN MATRICES*

STEFANO POZZA[†] AND VALERIA SIMONCINI[‡]

Abstract. The derivation of a-priori decay bounds for the entries of functions of banded matrices is of interest in a variety of applications. While decay bounds for functions of Hermitian banded matrices have been known for some time, the non-Hermitian case is an especially challenging setting. By using Faber polynomial series we explore the bounds obtainable by extending results for Hermitian matrices to banded non-Hermitian (not necessarily diagonalizable) matrices. Several special cases are treated, together with an application to the inexact Krylov approximation of matrix function evaluations. Numerical experiments illustrate the quality of all new bounds.

Key words. Banded matrices. Faber polynomials. Decay bounds. Matrix functions.

AMS subject classifications. 15A16, 65F50, 30E10, 65E05.

1. Introduction. Matrix functions have arisen as a reliable and a computationally attractive tool for solving a large variety of application problems; we refer the reader to [18] for a thorough discussion and references. The analysis of their properties and structure has recently attracted the interest of many practitioners. In particular, for a given square banded matrix A , the entries of the matrix function $f(A)$ for a sufficiently regular function f are characterized by a - typically exponential - decay pattern as they move away from the main diagonal. This phenomenon has been known for a long time, and it is at the basis of approximations and estimation strategies in many fields, from signal processing to quantum dynamics and multivariate statistics; see, e.g., [2, 3, 6] and their references. The interest in *a-priori* estimates that can accurately predict the decay rate of matrix functions has significantly grown in the past decades, and it has mainly focused on Hermitian matrices [13, 15, 25, 4, 33, 6, 11, 8]; the inverse and exponential functions have been given particular attention, due to their relevance in numerical analysis and other fields. Upper bounds usually take the form

$$|(f(A))_{k,\ell}| \leq c\rho^{|k-\ell|}, \quad (1.1)$$

where $\rho \in (0,1)$; both ρ and c depend on the spectral properties of A and on the domain of f , while ρ also strongly depends on the bandwidth of A .

The analysis of the decay pattern for banded *non-Hermitian* A is significantly harder, especially for non-normal matrices. In [5] Benzi and Razouk addressed this challenging case for diagonalizable matrices. They developed a bound of the type (1.1), where c also contains the eigenvector matrix condition number. In [22] the authors derive several qualitative bounds, mostly under the assumption that A is diagonally dominant. The exponential function provides a special setting, which has been explored in [20] and very recently by Wang in his PhD thesis [31]. In all these last

*Version of August 9, 2016. This work was supported in part by the FARB12SIMO grant, Università di Bologna and in part by INdAM-GNCS under the 2016 Project *Equazioni e funzioni di matrici con struttura: analisi e algoritmi*.

[†]Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, I-40127 Bologna, Italy (stefano.pozza@unibo.it)

[‡]Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, I-40127 Bologna, Italy (valeria.simoncini@unibo.it), and IMATI-CNR, Pavia.

three articles, and also in our approach, bounds on the decay pattern of banded non-Hermitian matrices are derived that avoid the explicit reference to the possibly large condition number of the eigenvector matrix. Specialized off-diagonal decay results have been obtained for certain normal matrices, see, e.g., [17, 11], and for analytic functions of banded matrices over C^* -algebras [2].

Starting with the pioneering work [12], most estimates for the decay behavior of the entries have relied on Chebyshev and Faber polynomials as technical tool, mainly for two reasons. Firstly, polynomials of banded matrices are again banded matrices, although the bandwidth increases with the polynomial degree. Secondly, sufficiently regular matrix functions can be written in terms of Chebyshev and Faber series, whose polynomial truncations enjoy nice approximation properties for a large class of matrices, from which an accurate description of the matrix function entries can be deduced.

We use Faber polynomials to obtain new bounds for functions that are analytic on the field of values of A , where A is a general non-Hermitian (not necessarily diagonalizable or diagonally dominant) matrix; see section 2 for the definition of field of values. The new estimates are able to capture the true decay pattern of matrix functions for a large class of functions, and can be combined with functions defined by the Laplace-Stieltjes transform. As an application, we consider the use of our results in the inexact Krylov approximation of matrix function evaluations; in particular, our new bounds can be used to devise a-priori relaxing thresholds for the inexact matrix-vector multiplications with A , whenever A is not available explicitly. These last results generalize a recently theory developed for $f(z) = z^{-1}$ and for the eigenvalue problem [29],[28]. Throughout the paper numerical experiments illustrate the quality of the new bounds.

The paper is organized as follows. Section 2 introduces some basic definitions and properties. In section 3 we use Faber polynomials to give a bound that can be adapted to approximate the entries of several matrix functions; as a sample we consider the functions e^A , $A^{-\frac{1}{2}}$ and $e^{-\sqrt{A}}$. Then we use the result for the exponential function to obtain bounds for functions defined by the Laplace-Stieltjes transform (subsection 3.1) and of Kronecker sums of banded matrices (subsection 3.2). In section 4 we first show that the new bounds can be used for a residual-type bound in the approximation of $f(A)\mathbf{v}$, for certain functions f by means of the Arnoldi algorithm. Then we describe how to employ this bound to reliably estimate the quality of the approximation when in the Arnoldi iteration the accuracy in the matrix-vector product is relaxed. We conclude with some remarks in section 5.

All our numerical experiments were performed using Matlab (R2013b) [23]. In all our experiments, the computation of the field of values employed the code in [9].

2. Preliminaries. We begin by recalling the definition of matrix function and some of its properties. Matrix functions can be defined in several ways (see [18, section 1]). For our presentation it is helpful to introduce the definition that employs the Cauchy integral formula.

DEFINITION 2.1. *Let $A \in \mathbb{C}^{n \times n}$ and f be an analytic function on some open $\Omega \subset \mathbb{C}$. Then*

$$f(A) = \int_{\Gamma} f(z) (zI - A)^{-1} dz,$$

with $\Gamma \subset \Omega$ a system of Jordan curves encircling each eigenvalue of A exactly once, with mathematical positive orientation.

When f is analytic Definition 2.1 is equivalent to other common definitions; see [26, section 2.3]. In what follows we will also consider certain matrix functions defined by integral measure transforms.

For $\mathbf{v} \in \mathbb{C}^n$ we denote with $\|\mathbf{v}\|$ the Euclidean vector norm, and for any matrix $A \in \mathbb{C}^{n \times n}$, with $\|A\|$ the induced matrix norm, that is $\|A\| = \sup_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|$. \mathbb{C}^+ denotes the open right-half complex plane. Moreover, we recall that the *field of values* of A is defined as the set $W(A) = \{\mathbf{v}^* A \mathbf{v} \mid \mathbf{v} \in \mathbb{C}^n, \|\mathbf{v}\| = 1\}$, where \mathbf{v}^* is the conjugate transpose of \mathbf{v} . We remark that the field of values of a matrix is a bounded convex subset of \mathbb{C} . In the following sections we need an approximation for the matrix norm of a matrix function. As proved by Crouzeix in [10], if A is a matrix with field of values $W(A)$, then for any function f in the Banach algebra of the functions analytic in the interior of $W(A)$ it holds

$$\|f(A)\| \leq C \sup_{w \in W(A)} |f(w)|, \quad (2.1)$$

with $C = 11.08$, conjecturing the stricter value $C = 2$. Notice that for some functions and some classes of matrices it does hold that $C = 2$. In the following we approximate the field of values of a matrix A with a subset $E \subset \mathbb{C}$, such that $W(A) \subseteq E$. Unless explicitly stated, E does not need to be symmetric with respect to the real axis. If E is a continuum (i.e., a non-empty, compact and connected subset of \mathbb{C}) with a connected complement, then by Riemann's mapping theorem there exists a function ϕ that maps the exterior of E conformally onto the exterior of the unitary disk $\{|z| \leq 1\}$. Subsets like E , the relative conformal maps ϕ , and their inverses ψ play a key role in the definition of Faber polynomials, which are a main tool in our analysis. For this reason from now on the notations E , ϕ , and ψ will be reserved to the objects defined above.

For the sake of simplicity and without loss of generality, our numerical experiments will mainly use Toeplitz matrices, which are constant along their diagonals. These matrices allow us to explore a large variety of spectral scenarios and non-normality properties, while providing a fully replicable experimental framework.

The (k, ℓ) element of a matrix A will be denoted by $(A)_{k,\ell}$. The set of banded matrices is defined as follows.

DEFINITION 2.2. *The notation $\mathcal{B}_n(\beta, \gamma)$ defines the set of banded matrices $A \in \mathbb{C}^{n \times n}$ with upper bandwidth $\beta \geq 0$ and lower bandwidth $\gamma \geq 0$, i.e., $(A)_{k,\ell} = 0$ for $\ell - k > \beta$ or $k - \ell > \gamma$.*

We observe that if $A \in \mathcal{B}_n(\beta, \gamma)$ with $\beta, \gamma \neq 0$, for

$$\xi := \begin{cases} \lceil (\ell - k)/\beta \rceil, & \text{if } k < \ell \\ \lceil (k - \ell)/\gamma \rceil, & \text{if } k \geq \ell \end{cases} \quad (2.2)$$

it holds that

$$(A^m)_{k,\ell} = 0, \quad \text{for every } m < \xi. \quad (2.3)$$

This characterization of banded matrices is a classical fundamental tool to prove the decay property of matrix functions, as sufficiently regular functions can be expanded in power series. Since we are interested in nontrivial banded matrices, in the following we shall assume that both β and γ are nonzero.

REMARK 2.3. *All our decay bounds describe the influence of the upper and lower bandwidths on the off-diagonal entry decay pattern. Numerical evidence indicates that the decay rate may also be influenced by the magnitude of the matrix nonzero entries,*

and in particular by the different magnitude of the elements in the upper and lower parts of the matrix. This property may determine a different decay rate for the two sides of the main diagonal, even for equal bandwidths β and γ . This asymmetry in the entry magnitude is partially accounted for by the shape of the field of values, and thus by our bounds. As a result, however, our estimates capture the slowest off-diagonal decay between the two sides.

Relation (2.3) can be extended to the case of a sparse, not necessarily banded, matrix A . Indeed, following the approach in [5] we can define the graph $G(A)$ describing the nonzero pattern of A , i.e., $G(A)$ is such that the vertex set of $G(A)$ consists of the indexes of the matrix $1, \dots, n$ and an edge (k, ℓ) is part of the graph if and only if $A_{k,\ell} \neq 0$. It thus follows that our analysis still holds if ξ is replaced by $d(k, \ell)$, the geodesic distance, i.e., the length of the shortest path between the nodes ℓ and k .

3. Decay bounds for analytic functions by Faber polynomials expansion. Faber polynomials extend the theory of power series to sets different from the disk, and can be effectively used to bound the entries of matrix functions.

Let E be a continuum with connected complement, and let us consider the relative conformal map ϕ satisfying the following conditions

$$\phi(\infty) = \infty, \quad \lim_{z \rightarrow \infty} \frac{\phi(z)}{z} = d > 0.$$

Hence, ϕ can be expressed by a Laurent expansion $\phi(z) = dz + a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \dots$. Furthermore, for every $n > 0$ we have

$$(\phi(z))^n = dz^n + a_{n-1}^{(n)} z^{n-1} + \dots + a_0^{(n)} + \frac{a_{-1}^{(n)}}{z} + \frac{a_{-2}^{(n)}}{z^2} + \dots$$

Then, the Faber polynomial for the domain E is defined by (see, e.g., [30])

$$\Phi_n(z) = dz^n + a_{n-1}^{(n)} z^{n-1} + \dots + a_0^{(n)}, \quad \text{for } n \geq 0.$$

If f is analytic on E then it can be expanded in a series of Faber polynomials for E , that is

$$f(z) = \sum_{j=0}^{\infty} f_j \Phi_j(z), \quad \text{for } z \in E;$$

[30, Theorem 2, p. 52]. Moreover, if the spectrum of A is contained in E and f is a function analytic in E , then the matrix function $f(A)$ can be expanded as follows (see, e.g., [30, p. 272])

$$f(A) = \sum_{j=0}^{\infty} f_j \Phi_j(A).$$

By properly extending the set E , this expansion allows us to establish a first intermediate result.

THEOREM 3.1. *Let $A \in \mathcal{B}_n(\beta, \gamma)$ with field of values contained in a convex continuum E with connected complement. Moreover, let $f(z) = \sum_{j=0}^{\infty} f_j \Phi_j(z)$ be the Faber expansion of f , in which Φ_j are Faber polynomials for E . Then*

$$|(f(A))_{k,\ell}| \leq 2 \sum_{j=\xi}^{\infty} |f_j|,$$

for $k \neq \ell$ and ξ defined by (2.2).

Proof. Let us consider super-diagonal elements of $(f(A))_{k,\ell}$ (for the sub-diagonal elements the proof is the same). Then $\mathbf{e}_k^T p_{m-1}(A) \mathbf{e}_\ell = 0$ for every polynomial p_{m-1} of degree at most $m-1$, with $m \geq |\ell - k|/\beta$. Hence, we get

$$f(A)_{k,\ell} = \sum_{j=0}^{\infty} f_j(\Phi_j(A))_{k,\ell} = \sum_{j=m}^{\infty} f_j(\Phi_j(A))_{k,\ell},$$

and thus $|f(A)_{k,\ell}| = |\mathbf{e}_k^T f(A) \mathbf{e}_\ell| \leq \sum_{j=m}^{\infty} |f_j| \|\Phi_j(A)\|$. Since E is convex, we conclude the proof using the inequality $\|\Phi_j(A)\| \leq 2$ proved in [1, Theorem 1]. \square

The approach used in the proof of Theorem 3.1 is novel and it is based on the expansion of $f(A)$ in a series of polynomials of A .

Notice that in [31, Theorem 3.8] a similar bound for the exponential function is derived in a different way. In [22] an analogous result is discussed, although our presentation is more complete and the proof different. By using Theorem 3.1 we can give general decay bounds for a large class of matrix functions.

THEOREM 3.2. *Let $A \in \mathcal{B}_n(\beta, \gamma)$ with field of values contained in a convex continuum E with connected complement whose boundary is Γ . Moreover, let ϕ be the conformal mapping of E , ψ be its inverse and $G_\tau = \{w : |\phi(w)| < \tau\}$. Let us assume that $\tau > 1$, f is analytic in G_τ and f is bounded on Γ_τ , the boundary of G_τ . Then*

$$\left| (f(A))_{k,\ell} \right| \leq 2 \frac{\tau}{\tau-1} \max_{|z|=\tau} |f(\psi(z))| \left(\frac{1}{\tau} \right)^\xi,$$

with ξ defined by (2.2).

Proof. From Theorem 3.1, $|(f(A))_{k,\ell}| \leq 2 \sum_{j=\xi}^{\infty} |f_j|$, where the Faber coefficients f_j are given by (see, e.g., [30, chapter III, Theorem 1])

$$f_j = \frac{1}{2\pi i} \int_{|z|=\tau} \frac{f(\psi(z))}{(z)^{j+1}} dz.$$

Hence, $|f_j| \leq \frac{1}{(\tau)^j} \max_{|z|=\tau} |f(\psi(z))|$. Therefore,

$$\begin{aligned} \left| (f(A))_{k,\ell} \right| &\leq 2 \max_{|z|=\tau} |f(\psi(z))| \sum_{j=\xi}^{\infty} \left(\frac{1}{\tau} \right)^j \\ &= 2 \max_{|z|=\tau} |f(\psi(z))| \left(\frac{1}{\tau} \right)^\xi \sum_{j=0}^{\infty} \left(\frac{1}{\tau} \right)^j = 2 \frac{\tau}{\tau-1} \max_{|z|=\tau} |f(\psi(z))| \left(\frac{1}{\tau} \right)^\xi. \quad \square \end{aligned}$$

We notice the similarity of this theorem with the one given in [16, Corollary 2.2] on the approximation error of analytic functions in terms of partial Faber series.

REMARK 3.3. *The bound of Theorem 3.2 has similarities with the one obtained by Benzi and Razouk in [5, Theorem 3.5]. Consider a diagonalizable matrix $A \in \mathcal{B}_n(\beta, \gamma)$ whose spectrum is contained in a convex continuum F with connected complement, and a function f analytic in $\{\psi(z) : |z| < \tau\}$, with $\tau > 1$ and ψ the inverse of the conformal map of F . Moreover let $\kappa(X) = \|X\| \|X^{-1}\|$ be the spectral condition*

number of the matrix X of eigenvectors of A . For a sufficiently large ξ and for every $\varepsilon > 0$ we can rewrite the bound in Theorem 3.5 of [5] as

$$\left| (f(A))_{k,\ell} \right| \lesssim \frac{3}{2} \kappa(X) \max_{|z| < R} |f(\psi(z))| \frac{1}{1 - (q + \varepsilon)} ((q + \varepsilon))^\xi, \quad (3.1)$$

with $q = \tau^{-1}$ and $R > (q + \varepsilon)^{-1}$. In this bound F needs to contain the spectrum of A , so it may be smaller than the set E we considered in Theorem 3.2 (which must contain $W(A)$). Hence, the value of τ may be allowed to be greater than the one in our bound. On the other hand, the bound (3.1) contains the factor $\kappa(X)$ which can be enormous. When A is a normal matrix, i.e., $AA^* = A^*A$, the two bounds have a similar rate. Indeed, in this case the convex hull of the spectrum is equal to the field of values and $\kappa(X) = 1$. However, in the non-normal case the two bounds can significantly differ. In particular, $\kappa(X)$ can be huge even when $W(A)$ is not much bigger than the spectrum. This can consistently be appreciated in our numerical experiments, where we report $\kappa(X)$ for completeness. For this reason, the new bound of Theorem 3.2 turns out to always be more descriptive than (3.1).

The choice of τ in Theorem 3.2, and thus the sharpness of the derived estimate, depends on the trade-off between the possible large size of f on the given region, and the exponential decay of $(1/\tau)^\xi$, and thus it produces an infinite family of bounds depending on the problem considered. As an example, we apply Theorem 3.2 to the approximation of three functions: $f(z) = e^z$, $f(z) = z^{-1/2}$ and $f(z) = e^{-\sqrt{z}}$, with z in a properly chosen domain.

COROLLARY 3.4. *Let $A \in \mathcal{B}_n(\beta, \gamma)$ with field of values contained in a closed set E whose boundary is a horizontal ellipse with semi-axes $a \geq b > 0$ and center $c = c_1 + ic_2 \in \mathbb{C}$, $c_1, c_2 \in \mathbb{R}$. Then*

$$\left| (e^A)_{k,\ell} \right| \leq 2e^{c_1} \frac{\xi + \sqrt{\xi^2 + a^2 - b^2}}{\xi + \sqrt{\xi^2 + a^2 - b^2} - (a + b)} \left(\frac{a + b}{\xi} \frac{e^{q(\xi)}}{1 + \sqrt{1 + (a^2 - b^2)/\xi^2}} \right)^\xi,$$

for $\xi > b$, with $q(\xi) = 1 + \frac{a^2 - b^2}{\xi^2 + \xi\sqrt{\xi^2 + a^2 - b^2}}$ and ξ as in (2.2).

Before we prove this result, we notice that for ξ large enough, the decay rate is of the form $((a + b)/(2\xi))^\xi$, that is, the decay is super-exponential.

Proof. Let $\rho = \sqrt{a^2 - b^2}$ be the distance between the foci and the center, and $R = (a + b)/\rho$. Then a conformal map for E is

$$\phi(w) = \frac{w - c - \sqrt{(w - c)^2 - \rho^2}}{\rho R},$$

and its inverse is

$$\psi(z) = \frac{\rho}{2} \left(Rz + \frac{1}{Rz} \right) + c, \quad (3.2)$$

see, e.g., [30, chapter II, Example 3]. Notice that

$$\max_{|z|=\tau} |e^{\psi(z)}| = \max_{|z|=\tau} e^{\Re(\psi(z))} = e^{\frac{\rho}{2} \left(R\tau + \frac{1}{R\tau} \right) + c_1}.$$

Hence by Theorem 3.2 we get

$$\left| (e^A)_{k,\ell} \right| \leq 2 \frac{\tau}{\tau - 1} e^{c_1} e^{\frac{\rho}{2} \left(R\tau + \frac{1}{R\tau} \right)} \left(\frac{1}{\tau} \right)^\xi.$$

The optimal value of $\tau > 1$ that minimizes $e^{\frac{\rho}{2}(R\tau + \frac{1}{R\tau})} \left(\frac{1}{\tau}\right)^\xi$ is

$$\tau = \frac{\xi + \sqrt{\xi^2 + \rho^2}}{\rho R}.$$

Moreover the condition $\tau > 1$ is satisfied if and only if $\xi > \frac{\rho}{2} \left(R - \frac{1}{R}\right) = b$. Finally, noticing that

$$\psi\left(\frac{\xi + \sqrt{\xi^2 + \rho^2}}{\rho R}\right) - c_1 = \frac{1}{2} \left(\xi + \sqrt{\xi^2 + \rho^2} + \frac{\rho^2}{\xi + \sqrt{\xi^2 + \rho^2}} \right) = \xi q(\xi),$$

and collecting ξ the proof is completed. \square

In the Hermitian, case, this bound is similar to bounds available in the literature. Indeed, let $A \in \mathcal{B}_n(\beta, \gamma)$ be Hermitian and such that $W(A) \subset [-2a, 0]$, $a > 0$. The following bound was obtained in [6],

$$\left| (e^A)_{k,\ell} \right| \leq 20 \frac{e^{-a/2}}{a} \left(\frac{ea}{2\xi} \right)^\xi, \quad \xi \geq a.$$

The bound in Corollary 3.4 has a similar decay rate. Indeed we can let $b \rightarrow 0$ in the bound, thus obtaining

$$\left| (e^A)_{k,\ell} \right| \leq 2 \frac{\xi + \sqrt{\xi^2 + a^2}}{\xi + \sqrt{\xi^2 + a^2} - a} e^{-a} \left(\frac{a}{\xi} \frac{e^{q(\xi)}}{1 + \sqrt{1 + a^2/\xi^2}} \right)^\xi, \quad \text{for } \xi > 0,$$

with $q(\xi) = 1 + \frac{a^2}{\xi^2 + \xi\sqrt{\xi^2 + a^2}}$. For $\xi \gg a$ the quantity in parentheses behaves like $(ea)/(2\xi)$. This is consistent with the fact that both estimates are based on Faber polynomial approximation.

EXAMPLE 3.5. *Figure 3.1 illustrates the quality of the bound in Corollary 3.4 for two different matrices. The top plots refer to $A \in \mathcal{B}_{200}(1, 1)$ with Toeplitz structure, $A = \text{Toeplitz}(-i, \underline{i}, -2)$, where the underlined element is on the diagonal, while the previous (resp. subsequent) values denote the lower (resp. upper) diagonal entries. The bottom plots refer to $A \in \mathcal{B}_{100}(2, 1)$, $A = \text{Toeplitz}(i, \underline{3i}, -i, -i)$. The left plots report the field of values of A (colored area), its eigenvalues (“ \times ”), and the ellipse used in the bound (dashed line). The right plots show the elements¹ of the t -th column of e^A (black solid line), and the corresponding bound from Corollary 3.4 (“ \times ”). In both examples the estimate is able to correctly capture the true (super-exponential) decay rate of the elements. For the two matrices, the condition number κ of the eigenvector matrix is approximately $\kappa = 4.0e + 29$ (top) and $\kappa = 5.5e + 13$ (bottom) (see Remark 3.3).*

Theorem 3.2 can be used to obtain bounds for many other matrix functions. An interesting example is the matrix inverse square root, which cannot be an analytic function in the whole complex plane. This property has crucial effects in the approximation, as the subsequent experiment shows.

COROLLARY 3.6. *Let $A \in \mathcal{B}_n(\beta, \gamma)$ with field of values contained in a closed set $E \subset \mathbb{C}^+$, whose boundary is a horizontal ellipse with semi-axes $a \geq b > 0$ and center*

¹The computation of e^A was performed with the Matlab function `expm`.

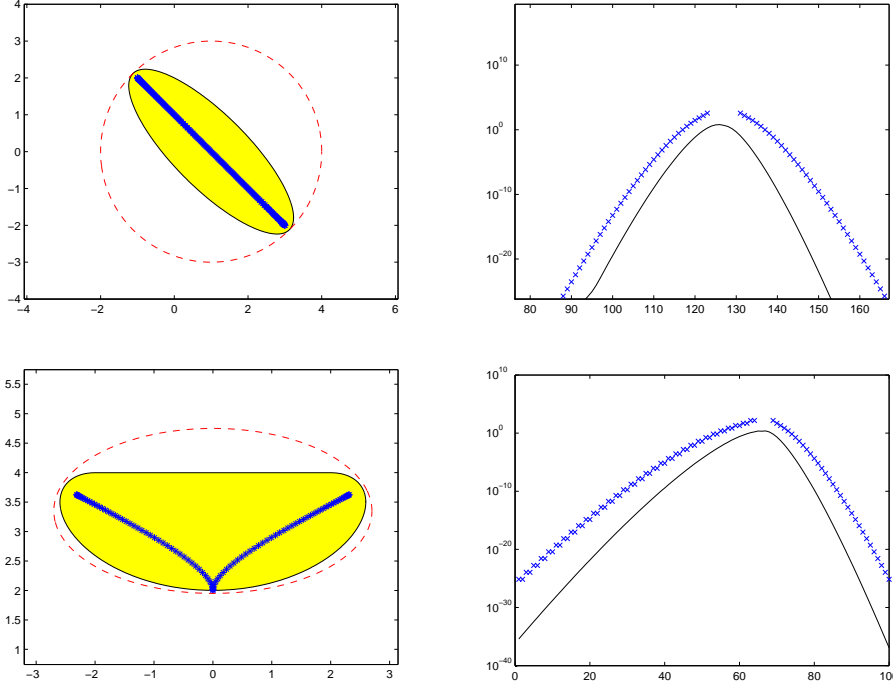


Fig. 3.1: Example 3.5. Left: field of values of A (colored region), its eigenvalues (“x”), and ellipse used in the bound (dashed line). Right (log-scale): the t -th column of e^A (solid line) and estimate of Corollary 3.4 (“x”). Top: $A = \text{Toeplitz}(-i, \underline{i}, -2) \in \mathbb{C}^{n \times n}$, $n = 200$, and $t = 127$. Bottom: $A = \text{Toeplitz}(i, \underline{3i}, -i, -i) \in \mathbb{C}^{n \times n}$, $n = 100$, and $t = 67$.

$c \in \mathbb{C}$. Then, for any $\varepsilon \in \mathbb{R}$ with $0 < \varepsilon \leq |c| - \sqrt{a(a+b)}$

$$\left| \left(A^{-\frac{1}{2}} \right)_{k,\ell} \right| \leq \frac{2}{\sqrt{\varepsilon}} q_2(a, b, c, \varepsilon) \left(\frac{a+b}{|c| - \varepsilon} \frac{1}{|1 + \sqrt{1 - (a^2 - b^2)/(c(1 - \varepsilon/|c|))^2}|} \right)^\xi,$$

with ξ defined by (2.2) and

$$q_2(a, b, c, \varepsilon) = \frac{\left| c(1 - \varepsilon/|c|) + \sqrt{c^2(1 - \varepsilon/|c|)^2 - (a^2 - b^2)^2} \right|}{\left| c(1 - \varepsilon/|c|) + \sqrt{c^2(1 - \varepsilon/|c|)^2 - (a^2 - b^2)^2} \right| - (a+b)}.$$

Proof. In order to uniquely define the square root of a matrix we consider the principal square root function, i.e., $\arg(\sqrt{z}) \in (-\pi/2, \pi/2]$ (see also [18, Chapter 1]). Thus the function $f(z) = (\sqrt{z})^{-1}$ is analytic in $\mathbb{C} \setminus (-\infty, 0]$.

The inverse of the conformal map ψ is given by (3.2). Hence, by Theorem 3.2 we can determine τ such that

$$\left| \left(A^{-\frac{1}{2}} \right)_{k,\ell} \right| \leq 2 \frac{\tau}{\tau - 1} \max_{|z|=\tau} \frac{1}{|\sqrt{\psi(z)}|} \left(\frac{1}{\tau} \right)^\xi.$$

Consider the circle with center the origin and radius $\varepsilon > 0$ and the ellipse $\{\psi(z), |z| = \bar{\tau}\}$ tangent to the circle and let $\varepsilon e^{i\varphi}$ be the tangent point between the two curves. Notice that since $\Re(c) > 0$ the ellipse is contained in $\mathbb{C} \setminus (-\infty, 0]$. If θ is such that $\bar{\tau} e^{i\theta} = \phi(\varepsilon e^{i\varphi})$, then $|\sqrt{\psi(\bar{\tau} e^{i\theta})}| = \sqrt{\varepsilon}$. Finally, $\bar{\tau} = |\phi(\varepsilon e^{i\varphi})|$, hence we can set

$$\tau = \min_{0 \leq \varphi \leq 2\pi} |\phi(\varepsilon e^{i\varphi})| = \left| \frac{c(1 - \varepsilon/|c|) + \sqrt{c^2(1 - \varepsilon/|c|)^2 - \rho^2}}{\rho R} \right|.$$

For the condition $\tau > 1$ to hold it must be

$$\left| (1 - \varepsilon/|c|) + \sqrt{(1 - \varepsilon/|c|)^2 - (\rho/c)^2} \right| \geq \rho R/|c|. \quad (3.3)$$

We then observe that for $\varepsilon \leq |c|$ and since $\Re\left(\sqrt{(1 - \varepsilon/|c|)^2 - (\rho/c)^2}\right) \geq 0$ we get

$$\begin{aligned} & \left| (1 - \varepsilon/|c|) + \sqrt{(1 - \varepsilon/|c|)^2 - (\rho/c)^2} \right|^2 \\ &= (1 - \varepsilon/|c|)^2 + \left| \sqrt{(1 - \varepsilon/|c|)^2 - (\rho/c)^2} \right|^2 + 2(1 - \varepsilon/|c|) \Re\left(\sqrt{(1 - \varepsilon/|c|)^2 - (\rho/c)^2}\right) \\ &\geq (1 - \varepsilon/|c|)^2 + \left| \sqrt{(1 - \varepsilon/|c|)^2 - (\rho/c)^2} \right|^2. \end{aligned}$$

Moreover, assuming $(1 - \varepsilon/|c|)^2 \geq (\rho/|c|)^2$, that is $|c| - \varepsilon \geq \rho$, it holds that

$$\left| \sqrt{(1 - \varepsilon/|c|)^2 - (\rho/c)^2} \right| \geq \left| \sqrt{(1 - \varepsilon/|c|)^2 - (\rho/|c|)^2} \right|.$$

The inequality

$$(1 - \varepsilon/|c|)^2 + \left| \sqrt{(1 - \varepsilon/|c|)^2 - (\rho/|c|)^2} \right|^2 \geq (\rho R/|c|)^2,$$

is satisfied for $\varepsilon \leq |c| - \sqrt{2\rho^2(1 + R^2)}/2 = |c| - \sqrt{a(a+b)}$, from which (3.3) follows. \square

For small $(a+b)/(|c| - \varepsilon)$, Corollary 3.6 predicts a linear asymptotic decay, in logarithmic scale, that goes like $\mathcal{O}((a+b)/(2|c|))^\xi$. The decay slope is well captured, while the actual rate may differ. In fact, the expression in Corollary 3.6 emphasizes the dependence of the new estimate on the function domain. The best possible τ is constrained by the condition that $\psi(z)$ with $|z| = \tau$ should lie inside a subspace in which f is analytic.

EXAMPLE 3.7. In Figure 3.2 we report on the quality of the bound of Corollary 3.6 for $f(z) = z^{-1/2}$, with $z \in \mathbb{C}^+$, and the two 100×100 matrices $A = \text{Toeplitz}(i, \underline{3i} + \underline{3}, -i, -i)$, and $A = \text{Toeplitz}(1, \underline{5}, 3)$; the eigenvector condition number is respectively $\kappa \approx 5.5e + 13, 1.2e + 24$. The contents of the plots are as in the previous example; here we report on the decay of the 67-th matrix function column. The function $A^{-1/2}$ was computed via the Matlab command `F=eye(n)/sqrtm(A)`. We set $\varepsilon = 0.05$. The field of values of the second matrix is an ellipse and can be sharply represented by the set E in Corollary 3.6. The obtained bound closely matches the true decay of the slowest decaying elements of the matrix. Also in the first example, however, the decay rate is captured almost correctly.

We conclude with the function $f(z) = e^{-\sqrt{z}}$, which will also be used in section 4.

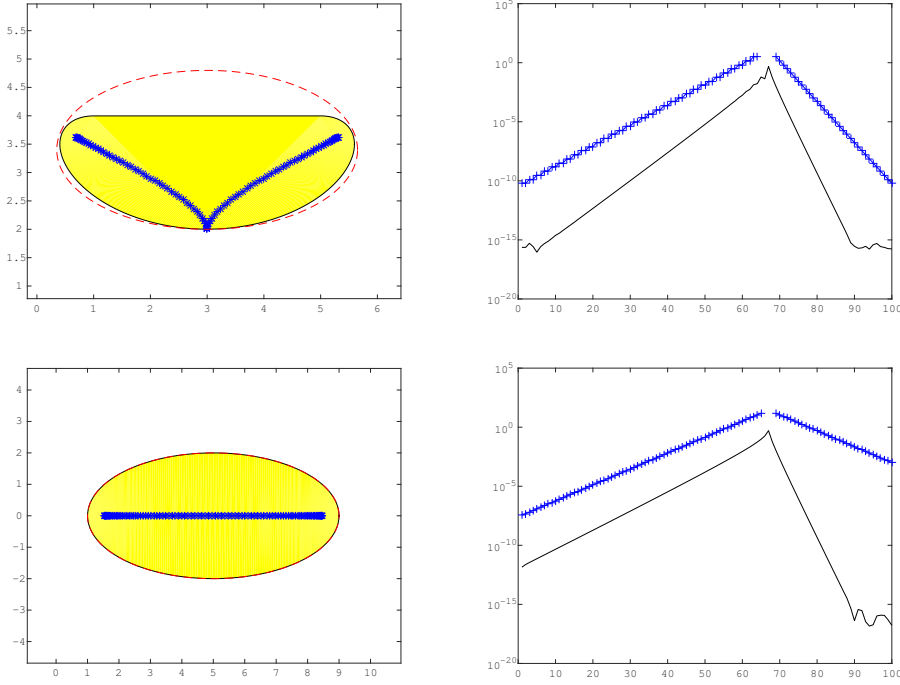


Fig. 3.2: Example 3.7. Top: matrix $A = \text{Toeplitz}(i, \underline{3} + 3i, -i, -i) \in \mathcal{B}_{100}(2, 1)$. Bottom: matrix $A = \text{Toeplitz}(1, \underline{5}, 3) \in \mathcal{B}_{100}(1, 1)$. Left: $W(A)$ (colored area), its eigenvalues (“x”), and ellipses used in the bound (dashed line). Right (log-scale): elements of the 67-th column of $A^{-\frac{1}{2}}$ (solid line) and bound from Corollary 3.6 (“x”).

COROLLARY 3.8. *Let $A \in \mathcal{B}_n(\beta, \gamma)$ with field of values contained in a closed set $E \subset \mathbb{C}^+$, whose boundary is a horizontal ellipse with semi-axes $a \geq b > 0$ and center $c \in \mathbb{C}$. Then,*

$$\left| \left(e^{-\sqrt{A}} \right)_{k,\ell} \right| \leq 2q_2(a, b, c, 0) \left(\frac{a+b}{|c|} \frac{1}{|1 + \sqrt{1 - (a^2 - b^2)/c^2}|} \right)^\xi,$$

with ξ defined by (2.2) and q_2 as in Corollary 3.6.

Proof. The function $f(z) = \exp(-\sqrt{z})$ is analytic in $\mathbb{C} \setminus (-\infty, 0)$. Notice that since we are considering the principal square root, then $\Re(\sqrt{z}) \geq 0$. Moreover, $|\exp(-\sqrt{z})| = \exp(-\Re(\sqrt{z})) \leq 1$. Hence, by Theorem 3.2 we can determine τ for which

$$\left| \left(e^{-\sqrt{A}} \right)_{k,\ell} \right| \leq 2 \frac{\tau}{\tau - 1} \left(\frac{1}{\tau} \right)^\xi.$$

We conclude the proof noticing that for

$$\tau = |\phi(0)| = \left| \frac{c + \sqrt{c^2 - \rho^2}}{\rho R} \right|$$

the ellipse $\{\psi(z), |z| = \tau\}$ is the maximal one contained in $\mathbb{C} \setminus (-\infty, 0)$. \square

REMARK 3.9. *For the sake of simplicity in the previous corollaries horizontal ellipses were employed. However, more general convex sets E may be considered. The previous bounds will change accordingly, since the optimal value for τ in Theorem 3.2 does depend on the parameters associated with E . For instance, for the exponential function and a vertical ellipse, we can derive the same bound as in Corollary 3.4 by letting $b > a$ (notice that this is different from exchanging the role of a and b in the bound). The proof of this fact is non-trivial but technical, and it is not reported.*

3.1. Bound for functions defined by the Laplace-Stieltjes transform.

Consider the nondecreasing measure $\mu(t)$ and the function defined by the Laplace-Stieltjes transform

$$f(z) = \int_0^\infty e^{-tz} d\mu(t), \quad (3.4)$$

which is convergent for $\Re(z) \geq 0$. Then, we can define the matrix function

$$f(A) = \int_0^\infty e^{-tA} d\mu(t) \quad (3.5)$$

for any matrix A having eigenvalues with positive real part. Typical examples are the inverse, which can be written as $z^{-1} = \int_0^\infty e^{-zt} d\mu_1(t)$ with $\mu_1(t) = t, t \geq 0$, and $(1 - e^{-z})/z = \int_0^\infty e^{-zt} d\mu_2(t)$, with $\mu_2(t) = t$, for $0 \leq t < 1$ and $\mu_2(t) = 1$ for $t \geq 1$.

The bound obtained in Corollary 3.4 for the exponential function can be used to derive new decay bounds for this class of matrix functions. To this end we follow the path proposed in the Hermitian case in [6, section 4.2]. For simplicity of exposition we consider the case when E is a disk. Nonetheless, the result can be easily generalized to an ellipse.

THEOREM 3.10. *Let f be as in (3.4) and let A be a banded matrix whose field of values is contained in a disk E with center c and radius R . Then*

$$\left| (f(A))_{k,\ell} \right| \leq 2\xi \int_0^{\frac{\xi}{R}} \frac{e^{-tc_1}}{\xi - Rt} \left(\frac{eRt}{\xi} \right)^\xi d\mu(t) + \int_{\frac{\xi}{R}}^\infty (e^{-tA})_{k,\ell} d\mu(t). \quad (3.6)$$

Proof. From Corollary 3.4 for the matrix $-tA$ we obtain

$$\left| (e^{-tA})_{k,\ell} \right| \leq 2 \frac{\xi}{\xi - Rt} e^{-c_1 t} \left(\frac{eRt}{\xi} \right)^\xi, \quad \text{for } \xi \geq Rt.$$

We conclude by using this relation in equation (3.5) for $t \leq \xi/R$. \square

Notice that the last term in (3.6) can be bounded by relation (2.1),

$$\int_{\frac{\xi}{R}}^\infty (e^{-tA})_{k,\ell} d\mu(t) \leq 11.08 \int_{\frac{\xi}{R}}^\infty \sup_{w \in W(A)} |\exp(-tw)| d\mu(t).$$

For the function $f(z) = (1 - e^{-z})/z$ the integral reduces to that on the interval $[0, 1]$ and the following bound can be obtained

$$\left| (A^{-1}(I - e^{-A}))_{k,\ell} \right| \leq 2\xi \int_0^1 \frac{e^{-tc_1}}{\xi - Rt} \left(\frac{eRt}{\xi} \right)^\xi dt, \quad \text{for } \xi > R. \quad (3.7)$$

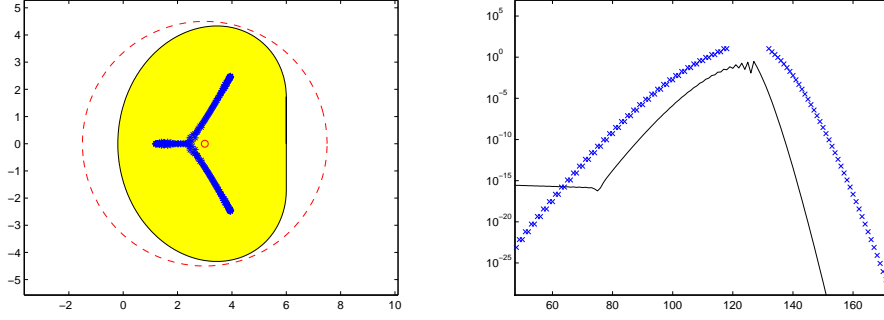


Fig. 3.3: Example 3.11. Matrix $A = \text{Toeplitz}(0.8, \underline{3}, -1, -3) \in \mathcal{B}_{200}(2, 1)$ and function $f(z) = (1 - e^{-z})/z$. Left: $W(A)$ (colored area), eigenvalues (“x”), and circle used in the bound (dashed line). Right (log-scale): entries of the 127-th column of $f(A)$ (solid line) and bound from Corollary 3.7 (“x”).

EXAMPLE 3.11. Figure 3.3 shows the behavior of the bound (3.7) for the matrix $A = \text{Toeplitz}(0.8, \underline{3}, -1, -3) \in \mathcal{B}_{200}(2, 1)$, with eigenvector condition number $\kappa = 3.3e + 43$, and the 127-th column of the matrix function $A^{-1}(I - e^{-A})$, computed in Matlab as $F = A \backslash (\text{eye}(n) - \text{expm}(-A))$. The integral was numerically estimated by the Matlab function `quadgk`. The description of the two plots is as in the previous examples. As for the other experiments associated with the exponential, the bound provides a quite good approximation of the true decay slope.

3.2. Bound for functions of Kronecker sums of matrices. As done in the recent literature (see, e.g., [6] and references therein), the peculiar oscillating decay of functions of Kronecker sums of banded matrices can be captured by exploiting the properties of the exponential function, when the Kronecker structure is present.

DEFINITION 3.12. Let A_1 and A_2 be two complex $n \times n$ matrices. The matrix $\mathcal{A} \in \mathbb{C}^{n^2 \times n^2}$ is the Kronecker sum of A_1 and A_2 if

$$\mathcal{A} = A_1 \oplus A_2 = A_1 \otimes I + I \otimes A_2.$$

The definition can be extended to three or more matrices, e.g.,

$$\mathcal{A} = A_1 \oplus A_2 \oplus A_3 = A_1 \otimes I \otimes I + I \otimes A_2 \otimes I + I \otimes I \otimes A_3.$$

The Kronecker sum of two matrices satisfies (see, e.g., [18, Theorem 10.9])

$$e^{A_1 \oplus A_2} = e^{A_1} \otimes e^{A_2}. \quad (3.8)$$

Functions of Kronecker sums of two banded matrices exhibit the typical decay away from the main diagonal, together with a refined decay associated with the bandwidth of the single matrices A_1, A_2 , giving rise to local “oscillations”. This behavior was characterized in [6, 8] for Hermitian positive definite matrices and a large class of functions. Thanks to the bounds in Theorem 3.10, we can generalize these results to non-Hermitian matrices, for matrix functions defined by (3.5).

It is useful to express the column and row indexes of an $n^2 \times n^2$ matrix $\mathcal{A} = A_1 \oplus A_2$ using the lexicographic ordering. Let $k = (k_1, k_2)$, $\ell = (\ell_1, \ell_2)$. Then $\mathcal{A}_{k,\ell}$

corresponds to the element in the $(k_2 - 1)n + k_1$ row and $(\ell_2 - 1)n + \ell_1$ column, with $k_1, k_2, \ell_1, \ell_2 \in \{1, \dots, n\}$. Therefore (see, e.g., [6, proof of Theorem 6.1])

$$(e^{A_1 \oplus A_2})_{k, \ell} = (e^{A_1} \otimes e^{A_2})_{k, \ell} = (e^{A_1})_{k_2, \ell_2} (e^{A_2})_{k_1, \ell_1}.$$

Let f be defined by (3.4), then

$$\begin{aligned} |(f(A_1 \oplus A_2))_{k, \ell}| &= \left| \int_0^\infty (e^{-t(A_1 \oplus A_2)})_{k, \ell} d\mu(t) \right| \\ &= \left| \int_0^\infty (e^{-tA_1})_{k_2, \ell_2} (e^{-tA_2})_{k_1, \ell_1} d\mu(t) \right| \\ &\leq \left(\int_0^\infty |(e^{-tA_1})_{k_2, \ell_2}|^2 d\mu(t) \right)^{\frac{1}{2}} \left(\int_0^\infty |(e^{-tA_2})_{k_1, \ell_1}|^2 d\mu(t) \right)^{\frac{1}{2}}. \end{aligned}$$

Let $A_j \in \mathcal{B}_n(\beta_j, \gamma_j)$. We define $\xi_j = \lceil (\ell_j - k_j) / \beta_j \rceil$ if $k_j < \ell_j$, or $\xi_j = \lceil (k_j - \ell_j) / \gamma_j \rceil$ if $k_j > \ell_j$, for $j = 1, 2$. Then, as in the proof of Theorem 3.10, we can bound the two last integrals as follows

$$\begin{aligned} \int_0^\infty |(e^{-tA_1})_{k_2, \ell_2}|^2 d\mu(t) &\leq \left(2\xi_2 \frac{(eR)^{\xi_2}}{(\xi_2)^{\xi_2}} \right)^2 \int_0^{\frac{\xi_2}{R}} \frac{e^{-2tc_1} t^{2\xi_2}}{(\xi_2 - Rt)^2} d\mu(t) + \\ &\quad + \int_{\frac{\xi_2}{R}}^\infty \left((e^{-tA_1})_{k_2, \ell_2} \right)^2 d\mu(t). \end{aligned}$$

As an example, consider the function $f(z) = (1 - e^{-z})/z$. From (3.7) we obtain

$$|(f(A \oplus A))_{k, \ell}| \leq 4 \frac{\xi_1 \xi_2 (eR)^{\xi_1 + \xi_2}}{(\xi_1)^{\xi_1} (\xi_2)^{\xi_2}} (I(\xi_1) I(\xi_2))^{\frac{1}{2}}, \quad \text{for } \xi_1, \xi_2 > R, \quad (3.9)$$

with

$$I(\xi) = \int_0^1 \frac{e^{-2tc_1} t^{2\xi}}{(\xi - Rt)^2} dt.$$

EXAMPLE 3.13. *Figure 3.4 illustrates the quality of the bound in (3.9) for $f(z) = (1 - e^{-z})/z$ and $\mathcal{A} = A \oplus A$. We consider $A = \text{Toeplitz}(-0.1, \underline{4}, 0.9i) \in \mathcal{B}_{30}(1, 1)$ ($\kappa = 7.8e + 13$), (top) and $A = \text{Toeplitz}(-1, \underline{4}, 1, 0.5) \in \mathcal{B}_{30}(2, 1)$ ($\kappa = 79.0$), (bottom), so that \mathcal{A} has dimension 900. The Matlab function `quadgk` was used to numerically evaluate the integral (3.9) for the two matrices. The matrix function $\mathcal{A}^{-1}(I - e^{-\mathcal{A}})$ was computed in Matlab as `F = A \setminus (\text{eye}(n) - \text{expm}(-A))`. The description of the two plots is as in the previous examples. The bound given by inequality (3.9) is able to predict the local and the global decay rate of the matrix function elements.*

4. Bounds for the approximation error of exact and inexact Arnoldi methods. Given a matrix $A \in \mathbb{C}^{n \times n}$ and a vector $\mathbf{v} \in \mathbb{C}^n$, we define the m th Krylov subspace generated by A and \mathbf{v} as

$$\mathcal{K}_m(A, \mathbf{v}) = \text{span}\{\mathbf{v}, A\mathbf{v}, \dots, A^{m-1}\mathbf{v}\}.$$

For $\mathbf{v}_1 = \mathbf{v}/\|\mathbf{v}\|$ and $m \geq 1$, the m th step of the Arnoldi algorithm determines an orthonormal basis $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ for $\mathcal{K}_m(A, \mathbf{v})$, the subsequent orthonormal basis vector \mathbf{v}_{m+1} , an $m \times m$ upper Hessenberg matrix H_m , and a scalar $h_{m+1, m}$ such that

$$A\mathbf{v}_m = V_m H_m + h_{m+1, m} \mathbf{v}_{m+1} \mathbf{e}_m^T,$$

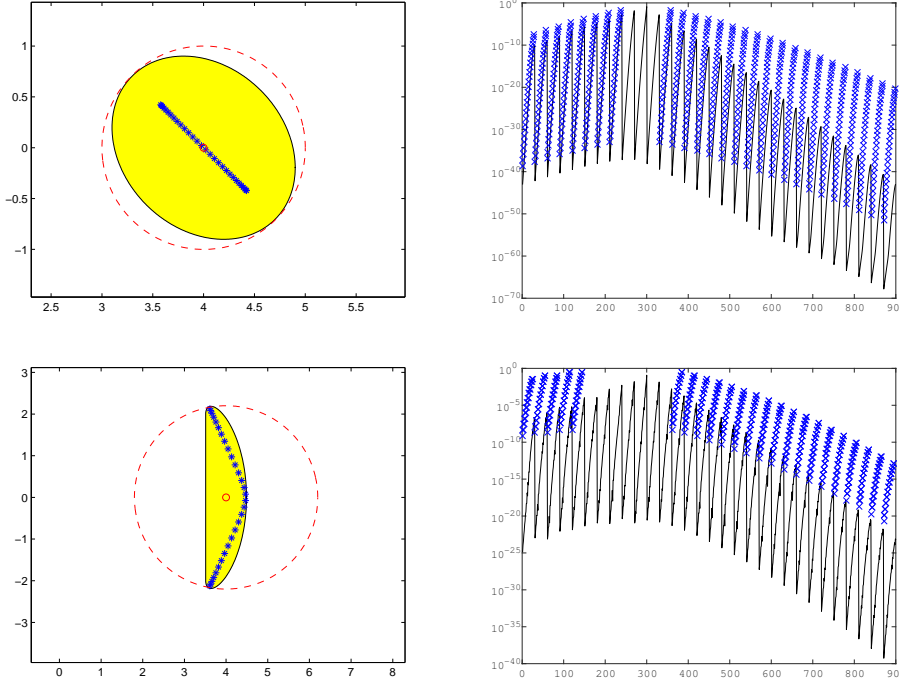


Fig. 3.4: Example 3.13. Matrix $\mathcal{A} = A \oplus A$. Top: $A = \text{Toeplitz}(-0.1, \underline{4}, 0.9i) \in \mathcal{B}_{30}(1, 1)$. Bottom: $A = \text{Toeplitz}(-1, \underline{4}, 1, 0.5) \in \mathcal{B}_{30}(2, 1)$. Left: $W(A)$ (colored red), eigenvalues (“x”), and ellipse used in the bound (dashed line). Right (log-scale): the 300-th column of $(\mathcal{A})^{-1}(I - e^{-\mathcal{A}})$ (solid line) and the values of the equation (3.9) bound (“x”).

where $V_m = [\mathbf{v}_1, \dots, \mathbf{v}_m]$. Due to the orthogonality of the columns of $[V_m, v_{m+1}]$, the matrix H_m is the projection and restriction of A onto $\mathcal{K}_m(A, \mathbf{v})$, that is $H_m = V_m^* A V_m$. The Arnoldi approximation to $f(A)\mathbf{v}$ is given as $V_m f(H_m)\mathbf{e}_1$; see, e.g., [18, Ch.13]. The case of the matrix exponential has been especially considered. Estimates of the error norm $\|e^{-tA}\mathbf{v} - V_m e^{-tH_m}\mathbf{e}_1\|$ for A non-normal have been given for instance by Saad [27], by Lubich and Hochbruck in [19], and recently by Wang and Ye in [32] and [31].

In the Hermitian case, bounds of the Arnoldi approximation have been used to obtain upper estimates for the entries decay; see for instance [6] for the exponential function. With our new results we can again exploit this connection but in the reverse direction. More precisely, let A be a complex $n \times n$ matrix and \mathbf{v} be a unit norm vector. By using decay bounds for the entries of $f(H_m)\mathbf{e}_1$ with H_m upper Hessenberg, we next show that we can give a bound for a specifically defined residual associated with the approximation of $f(A)\mathbf{v}$ in the case of generic non-Hermitian A and several different functions; these bounds complement those available in the already mentioned literature for the Arnoldi approximation. The quantity $|\mathbf{e}_m^T f(H_m)\mathbf{e}_1|$ is commonly used to monitor the accuracy of the approximation $\|f(A)\mathbf{v} - V_m f(H_m)\mathbf{e}_1\|$. Notice that $|\mathbf{e}_m^T f(H_m)\mathbf{e}_1| = |(f(H_m))_{m,1}|$, the last entry of the first column of $f(H_m)$. In

the case of the exponential, $e^{-tA}\mathbf{v}$, the quantity $r_m(t) = |h_{m+1,m}\mathbf{e}_m^T e^{-tH_m}\mathbf{e}_1|$ can be interpreted as the “residual” norm of an associated differential equation, see [7] and references therein; this is true also for other functions, see, e.g., [14, section 6]. Indeed, assume that $\mathbf{y}(t) = f(tA)\mathbf{v}$ is the solution to the differential equation $y^{(d)} = Ay$ for some d th derivative, $d \in \mathbb{N}$ and specified initial conditions for $t = 0$. Let $\mathbf{y}_m(t) = V_m f(tH_m)\mathbf{e}_1 =: V_m \hat{\mathbf{y}}_m(t)$. The vector $\hat{\mathbf{y}}_m(t)$ is the solution to the projected equation $\hat{\mathbf{y}}_m^{(d)} = H_m \hat{\mathbf{y}}_m$ with initial condition $\hat{\mathbf{y}}_m(0) = \mathbf{e}_1$. The differential equation residual $\mathbf{r}_m = A\mathbf{y}_m - \mathbf{y}_m^{(d)}$ can be used to monitor the accuracy of the approximate solution. Indeed, using the definition of \mathbf{y}_m and the Arnoldi relation,

$$\begin{aligned} \mathbf{r}_m &= A\mathbf{y}_m - \mathbf{y}_m^{(d)} = AV_m f(tH_m)\mathbf{e}_1 - \mathbf{y}_m^{(d)} \\ &= V_m H_m f(tH_m)\mathbf{e}_1 - V_m (f(tH_m))^{(d)}\mathbf{e}_1 + \mathbf{v}_{m+1} h_{m+1,m} \mathbf{e}_m^T f(tH_m)\mathbf{e}_1 \\ &= V_m (H_m \hat{\mathbf{y}}_m - \hat{\mathbf{y}}_m^{(d)}) + \mathbf{v}_{m+1} h_{m+1,m} \mathbf{e}_m^T f(tH_m)\mathbf{e}_1 \\ &= \mathbf{v}_{m+1} h_{m+1,m} \mathbf{e}_m^T f(tH_m)\mathbf{e}_1. \end{aligned}$$

Without loss of generality in the following we consider $t = 1$. We remark that the property $H_m = V_m^* A V_m$ ensures that the field of values of H_m is contained in that of A , so that our theory can be applied using A as reference matrix to individuate the spectral region of interest. Let a, b be the semi-axes and $c = c_1 + ic_2$ the center of an elliptical region E containing the field of values of A and $\xi = m - 1$. From Corollary 3.4 for $m > b + 1$ we get the inequality

$$|r_m(1)| \leq h_{m+1,m} 2e^{-c_1} p(m) \left(\frac{e^{q(m-1)}(a+b)}{m-1 + \sqrt{(m-1)^2 + (a^2 - b^2)}} \right)^{m-1}, \quad (4.1)$$

with

$$q(m-1) = 1 + \frac{(a^2 - b^2)}{(m-1)^2 + (m-1)\sqrt{(m-1)^2 + (a^2 - b^2)}}$$

and

$$p(m) = \frac{m-1 + \sqrt{(m-1)^2 + (a^2 - b^2)}}{m-1 + \sqrt{(m-1)^2 + (a^2 - b^2)} - (a+b)}.$$

In [31, 32] a similar bound is proposed, where however a continuum E with rectangular shape is considered, instead of the elliptical one we take in Corollary 3.4. Experiments suggest that the sharpness of these bounds depends on which set E better approximates the matrix field of values.

EXAMPLE 4.1. Figure 4.1 shows the behavior of the bound in (4.1) for the residual of the Arnoldi approximation of $e^{-A}\mathbf{v}$, with $A = \text{Toeplitz}(1, \underline{2}, 0.1, -1) \in \mathcal{B}_{200}(2, 1)$ (top), A the matrix pde225 of the Matrix Market repository [24] (bottom) and $\mathbf{v} = (1, \dots, 1)^T / \sqrt{n}$. The left figure shows the field of values of the matrix A (yellow area), its eigenvalues (blue crosses), and the horizontal ellipse used in the bound (red dashed line). On the right we plot the residual associated with the Arnoldi approximation as the iteration proceeds (black solid line), and the corresponding values of the bound (blue crosses). Matrix exponentials were computed by the `expm` Matlab function.

In an inexact Arnoldi procedure A is not known exactly. This may be due for instance to the fact that A is only implicitly available via functional operations with

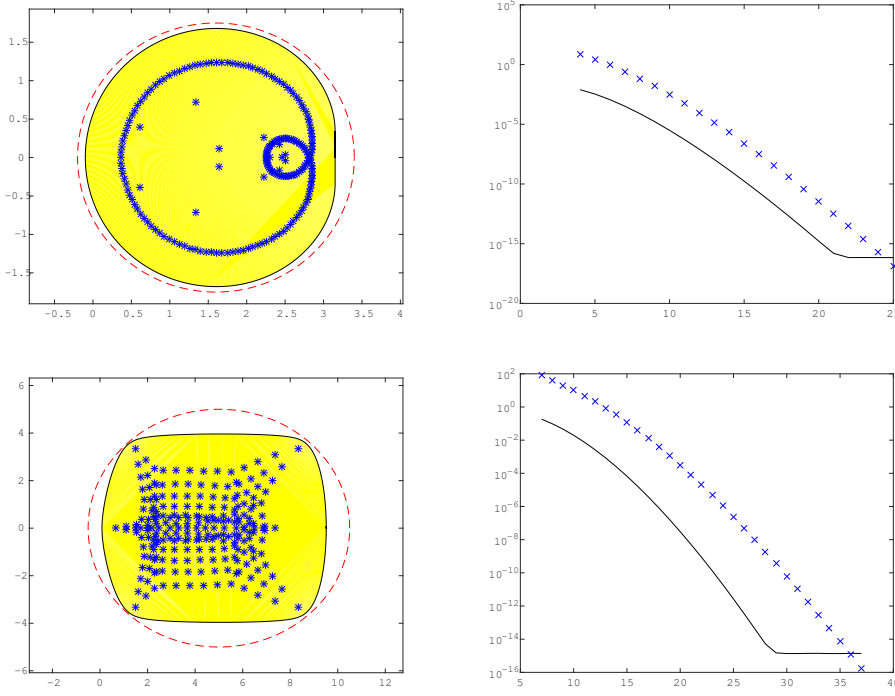


Fig. 4.1: Example 4.1. Approximation of $e^{-A}\mathbf{v}$, with $\mathbf{v} = (1, \dots, 1)^T / \sqrt{n}$. Top: $A = \text{Toeplitz}(1, \underline{2}, 0.1, -1) \in \mathcal{B}_{200}(2, 1)$. Bottom: matrix `pde225`. Left: $W(A)$ (yellow area), eigenvalues of A (blue crosses), and enclosing ellipse E (red dashed line). Right: residual norm as the Arnoldi iteration proceeds in the approximation (black solid line), and residual bound in (4.1) (blue crosses).

a vector, which can be approximated at some accuracy. To proceed with our analysis we can formalize this inexactness at each iteration k as

$$\tilde{\mathbf{v}}_{k+1} = A\mathbf{v}_k + \mathbf{w}_k \approx A\mathbf{v}_k. \quad (4.2)$$

Typically, some form of accuracy criterion is implemented, so that $\|\mathbf{w}_k\| < \epsilon$ for some ϵ . It may be that a different value of this tolerance is used at each iteration k , so that $\epsilon = \epsilon_k$. The new vector $\tilde{\mathbf{v}}_{k+1}$ is then orthonormalized with respect to the previous basis vectors to obtain \mathbf{v}_{k+1} . In compact form, the original Arnoldi relation becomes

$$(A + \mathcal{E}_m)V_m = V_m H_m + h_{m+1,m} \mathbf{v}_{m+1} \mathbf{e}_m^T, \quad \mathcal{E}_m = [\mathbf{w}_1, \dots, \mathbf{w}_m] V_m^*.$$

Here H_m is again upper Hessenberg, however, $H_m = V_m^*(A + \mathcal{E}_m)V_m$. Moreover, \mathcal{E}_m changes as m grows. The differential equation residual can be defined in the same way as for the exact case, $\mathbf{r}_m = A\mathbf{y}_m - \mathbf{y}_m^{(d)}$, however the inexact Arnoldi relation should be considered to proceed further. Indeed,

$$\begin{aligned} \mathbf{r}_m &= A\mathbf{y}_m - \mathbf{y}_m^{(d)} = AV_m f(tH_m) \mathbf{e}_1 - \mathbf{y}_m^{(d)} \\ &= -\mathcal{E}_m V_m f(tH_m) \mathbf{e}_1 + V_m H_m f(tH_m) \mathbf{e}_1 - \mathbf{y}_m^{(d)} + \mathbf{v}_{m+1} h_{m+1,m} \mathbf{e}_m^T f(tH_m) \mathbf{e}_1 \\ &= -[\mathbf{w}_1, \dots, \mathbf{w}_m] f(tH_m) \mathbf{e}_1 + \mathbf{v}_{m+1} h_{m+1,m} \mathbf{e}_m^T f(tH_m) \mathbf{e}_1. \end{aligned}$$

Note that $\|\mathbf{r}_m\|$ is not available, since A cannot be applied exactly. However, with the previous notation we can write $\|\mathbf{r}_m\| \leq ||\|\mathbf{r}_m\| - r_m| + r_m$ where

$$||\|\mathbf{r}_m\| - r_m| \leq \|[\mathbf{w}_1, \dots, \mathbf{w}_m]f(tH_m)\mathbf{e}_1\|.$$

Therefore, checking the available r_m provides a good measure of the accuracy in the function estimation as long as $\|[\mathbf{w}_1, \dots, \mathbf{w}_m]f(tH_m)\mathbf{e}_1\|$ is smaller than the requested tolerance for the final accuracy of the computation.

Clearly, $\|[\mathbf{w}_1, \dots, \mathbf{w}_m]f(tH_m)\mathbf{e}_1\| \leq \|[\mathbf{w}_1, \dots, \mathbf{w}_m]\| \|f(tH_m)\mathbf{e}_1\|$ so that the criterion $\|\mathbf{w}_k\| < \epsilon$ can be used to monitor the quality of the approximation to $f(A)\mathbf{v}$ by means of r_m . However, a less stringent criterion can be devised. Following similar discussions in [29],[28], we write

$$\|[\mathbf{w}_1, \dots, \mathbf{w}_m]f(tH_m)\mathbf{e}_1\| = \left\| \sum_{j=1}^m \mathbf{w}_j \mathbf{e}_j^T f(tH_m)\mathbf{e}_1 \right\| \leq \sum_{j=1}^m \|\mathbf{w}_j\| |\mathbf{e}_j^T f(tH_m)\mathbf{e}_1|, \quad (4.3)$$

where we assume that $\|\mathbf{w}_j\| < \epsilon_j$, that is the accuracy in the computation with A varies with j . Hence, $\|[\mathbf{w}_1, \dots, \mathbf{w}_m]f(tH_m)\mathbf{e}_1\|$ is small when either $\|\mathbf{w}_j\|$ or $|\mathbf{e}_j^T f(tH_m)\mathbf{e}_1|$ is small, and not necessarily both. By exploiting the exponential decay of the entries of $f(tH_m)\mathbf{e}_1$, we can infer that $\|\mathbf{w}_j\|$ is in fact allowed to grow with j , according with the exponential decay of the corresponding entries of $f(tH_m)\mathbf{e}_1$, without affecting the overall accuracy. A-priori bounds on $|\mathbf{e}_j^T f(tH_m)\mathbf{e}_1|$ can be used to select ϵ_j when estimating $A\mathbf{v}_j$. This relaxed strategy can significantly decrease the computational cost of matrix function evaluations whenever applying A accurately is expensive. However, notice that the field of values of H_m is contained in the field of values of $A + \mathcal{E}_m$. Hence if $W(A)$ is contained in an ellipse ∂E of semi-axes a, b and center c then $W(A + \mathcal{E}_m) \subset W(A) + W(\mathcal{E}_m)$. Since

$$\sup_{\|z\|=1} |z^* \mathcal{E}_m z| \leq \sup_{\|z\|=1} \|\mathcal{E}_m z\| \leq \sqrt{\sum_{j=1}^m \|\mathbf{w}_j\|^2} \leq \sqrt{\sum_{j=1}^m \epsilon_j^2} =: \epsilon^{(m)},$$

the set $W(\mathcal{E}_m)$ is contained in the disk centered at the origin and radius $\epsilon^{(m)}$. Therefore, $W(A) + W(\mathcal{E}_m)$ is contained in any set whose boundary has minimal distance from ∂E not smaller than $\epsilon^{(m)}$. One such set is contained in the ellipse ∂E_m with semi-axes $a(1 + \epsilon^{(m)}/b)$, $b + \epsilon^{(m)}$ and center c . Indeed, $z \in \partial E_m$ can be parameterized as

$$z = \left(1 + \frac{\epsilon^{(m)}}{b}\right) \frac{\rho}{2} \left(Re^{i\theta} + \frac{1}{Re^{i\theta}}\right) + c, \quad 0 \leq \theta \leq 2\pi,$$

with $\rho = \sqrt{a^2 - b^2}$, $R = (a + b)/\rho$. The distance between z and the ellipse ∂E is

$$\left| \frac{\epsilon^{(m)}}{b} \frac{\rho}{2} \left(Re^{i\theta} + \frac{1}{Re^{i\theta}}\right) \right| \geq \left| \frac{\epsilon^{(m)}}{b} \frac{\rho}{2} \left(R - \frac{1}{R}\right) \right| = \epsilon^{(m)}.$$

Let us fix a tolerance $tol > 0$, a maximum number of iterations m and a value $\epsilon^{(m)} > 0$. Moreover, let s_j be the upper bound for $|\mathbf{e}_j^T f(H_m)\mathbf{e}_1|$ from Corollary 3.4, with the above ellipse ∂E_m . The following choice for the accuracy for $j = 1, \dots, m$

$$\bar{\epsilon}_j = \begin{cases} \frac{tol}{m} \max(1, \frac{1}{s_j}), & \text{if } \frac{tol}{m s_j} < \frac{\sqrt{(\epsilon^{(m)})^2 - \sum_{k=1}^{j-1} \epsilon_k^2}}{m-j+1} \\ \frac{\sqrt{(\epsilon^{(m)})^2 - \sum_{k=1}^{j-1} \epsilon_k^2}}{m-j+1}, & \text{otherwise} \end{cases} \quad (4.4)$$

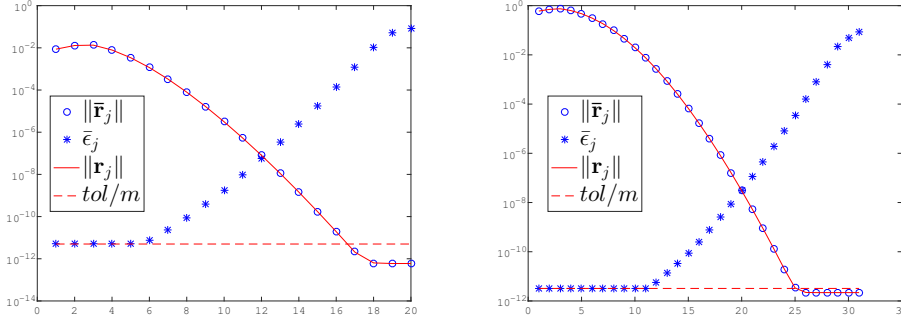


Fig. 4.2: Example 4.2, approximation of $e^{-A}\mathbf{v}$ with $\mathbf{v} = (1, \dots, 1)^T/\sqrt{n}$. Residual norm $\|\mathbf{r}_j\|$ with constant accuracy $\epsilon_j = \text{tol}/m$, and residual norm $\|\bar{\mathbf{r}}_j\|$ with $\epsilon_j = \bar{\epsilon}_j$ by (4.4) as the inexact Arnoldi method proceeds. Left: For $A = \text{Toeplitz}(1, \underline{2}, 0.1, -1) \in \mathcal{B}_{200}(1, 1)$. Right: For matrix **pde225** from the Matrix Market repository [24].

gives $\sqrt{\sum_{j=1}^m \bar{\epsilon}_j^2} \leq \epsilon^{(m)}$ and $|\|\mathbf{r}_m\| - r_m| \leq \text{tol}$.

EXAMPLE 4.2. We consider the inexact Arnoldi procedure for the approximation of $\exp(-A)\mathbf{v}$, so that the norm of the differential equation residual is lower than a tolerance tol . The inexact matrix-vector product was implemented as in (4.2), where \mathbf{w}_j is a random vector of norm ϵ_j .

Figure 4.2 reports our results for $\mathbf{v} = (1, \dots, 1)^T/\sqrt{n}$ and the same matrices as in Example 4.1: $A = \text{Toeplitz}(1, \underline{2}, 0.1, -1) \in \mathcal{B}_{200}(2, 1)$ (left), and the matrix **pde225** from the Matrix Market repository [24] (right). For constant accuracy $\epsilon_j = \text{tol}/m$ (dashed line), the solid line shows the residual norm $\|\mathbf{r}_j\|$ as the iteration j proceeds. For variable accuracy $\epsilon_j = \bar{\epsilon}_j$ obtained from (4.4) (stars) the circles display the residual norm $\|\bar{\mathbf{r}}_j\|$. We set $\text{tol} = 10^{-10}$ and $\epsilon^{(m)} = 10^{-1}$. The maximum approximation space dimension m was chosen as the smallest value for which the bound (4.1) is lower than tol , respectively $m = 20$ and $m = 31$. The fields of values of the matrices can be obtained starting from those reported in the left plots of Figure 4.1, where however now the original semi-axes a, b of the elliptical sets considered for the computation of s_j are increased by $\epsilon^{(m)}/b$ and $\epsilon^{(m)}$ respectively. The plots show visually overlapping residual norm histories for the two choices of ϵ_j , illustrating that in practice no loss of information takes place during the relaxing strategy.

Consider the differential equation $y^{(2)} = Ay$, with $y(0) = \mathbf{v}$. Its solution can be expressed as $y(t) = \exp(-t\sqrt{A})\mathbf{v}$, and our results can be applied to this case as well. This time the upper bound s_j for $|\mathbf{e}_m^T f(H_m) \mathbf{e}_1|$ is obtained from Corollary 3.8.

EXAMPLE 4.3. For the same experimental setting as in Example 4.2 we consider approximating $\exp(-\sqrt{A})\mathbf{v}$, for $A = \text{Toeplitz}(-1, 1, \underline{3}, 0.1) \in \mathcal{B}_{200}(1, 2)$, $\mathbf{v} = (1, \dots, 1)^T/\sqrt{200}$ and $m = 35$. Figure 4.3 reports on our findings, with the same description as for the previous example. Here s_j in (4.4) is obtained from Corollary 3.8, and it is used to relax the accuracy ϵ_j . Similar considerations apply.

5. Conclusions. We have proved that for a large class of functions sharp bounds on the off-diagonal decay pattern of functions of non-normal matrices can be obtained. Different proof strategies have been adopted, to comply with the analyticity properties

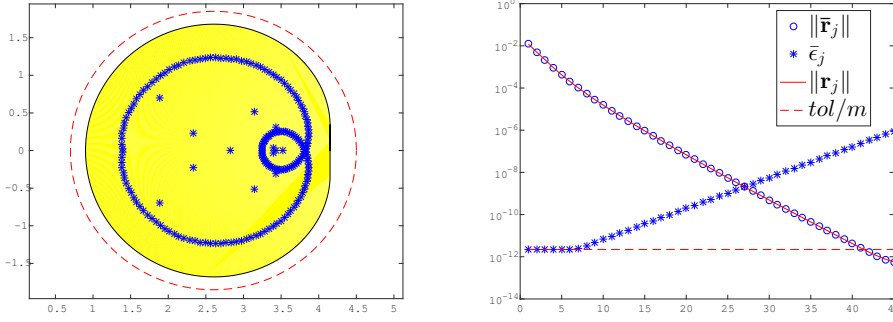


Fig. 4.3: Example 4.3. Approximation of $e^{-A}\mathbf{v}$ with $A = \text{Toeplitz}(-1, 1, \underline{3}, 0.1) \in \mathcal{B}_{200}(1, 2)$ and $\mathbf{v} = (1, \dots, 1)^T/\sqrt{n}$. Left: spectral information. Right: Residual norm $\|\mathbf{r}_j\|$ with constant accuracy $\epsilon_j = \text{tol}/m$, and residual norm $\|\bar{\mathbf{r}}_j\|$ with $\epsilon_j = \bar{\epsilon}_j$ by (4.4) as the inexact Arnoldi method proceeds.

of the considered functions, and the spectral properties of the given matrices. As expected, our bounds are also influenced by the dependence between the predicted decay rate and the shape and dimension of the set enclosing the field of values of A . The closer E is to the field of values, the sharper the bound. We have shown that our decay estimates can be used in monitoring the inexactness of matrix-vector products in Arnoldi approximations of matrix functions applied to a vector. Similar results can be obtained for other Krylov-type approximations whose projection and restriction matrix H_m has a semi-banded structure. This is the case for instance of the Extended Krylov subspace approximation; see, e.g., [21] and references therein.

Acknowledgements. We are indebted with Leonid Knizhnerman for a careful reading of a previous version of this manuscript, and for his many insightful remarks which led to great improvements of our results. We also thank Michele Benzi for several suggestions.

REFERENCES

- [1] BERNHARD BECKERMANN, *Image numérique, GMRES et polynômes de Faber*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 855–860.
- [2] MICHELE BENZI AND PAOLA BOITO, *Decay properties for functions of matrices over C^* -algebras*, Linear Algebra Appl., 456 (2014), pp. 174–198.
- [3] MICHELE BENZI, PAOLA BOITO, AND NADER RAZOUK, *Decay properties of spectral projectors with applications to electronic structure*, SIAM Rev., 55 (2013), pp. 3–64.
- [4] MICHELE BENZI AND GENE H. GOLUB, *Bounds for the entries of matrix functions with applications to preconditioning*, BIT, 39 (1999), pp. 417–438.
- [5] MICHELE BENZI AND NADER RAZOUK, *Decay bounds and $O(n)$ algorithms for approximating functions of sparse matrices*, Electron. Trans. Numer. Anal., 28 (2007), pp. 16–39.
- [6] MICHELE BENZI AND VALERIA SIMONCINI, *Decay bounds for functions of Hermitian matrices with banded or Kronecker structure*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1263–1282.
- [7] MIKE A. BOTCHEV, VOLKER GRIMM, AND MARLIS HOCHBRUCK, *Residual, restarting and Richardson iteration for the matrix exponential*, SIAM J. Sci. Comput., 35 (2013), pp. A1376–A1397.
- [8] CLAUDIO CANUTO, VALERIA SIMONCINI, AND MARCO VERANI, *On the decay of the inverse of matrices that are sum of Kronecker products*, Linear Algebra Appl., 452 (2014), pp. 21–39.

- [9] CARL C. COWEN AND ELAD HAREL, *An effective algorithm for computing the numerical range*, <https://www.math.iupui.edu/~ccowen/Downloads/33NumRange.html>, 1995.
- [10] MICHEL CROUZEX, *Numerical range and functional calculus in Hilbert space*, J. Funct. Anal., 244 (2007), pp. 668–690.
- [11] NICOLETTA DEL BUONO, LUCIANO LOPEZ, AND ROBERTO PELUSO, *Computation of the exponential of large sparse skew-symmetric matrices*, SIAM J. Sci. Comput., 27 (2005), pp. 278–293.
- [12] STEPHEN DEMKO, WILLIAM F. MOSS, AND PHILIP W. SMITH, *Decay rates for inverses of band matrices*, Math. Comp., 43 (1984), pp. 491–499.
- [13] STEPHEN G. DEMKO, *Inverses of band matrices and local convergence of spline projections*, SIAM J. Numer. Anal., 14 (1977), pp. 616–619.
- [14] VLADIMIR DRUSKIN AND LEONID KNIZHNERMAN, *Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic*, Numer. Linear Algebra Appl., 2 (1995), pp. 205–217.
- [15] VICTOR ELJKHOUT AND BEN POLMAN, *Decay rates of inverses of banded M -matrices that are near to Toeplitz matrices*, Linear Algebra Appl., 109 (1988), pp. 247–277.
- [16] STEPHEN W. ELLACOTT, *Computation of Faber series with application to numerical polynomial approximation in the complex plane*, Math. Comp., 40 (1983), pp. 575–587.
- [17] ROLAND FREUND, *On polynomial approximations to $f_a(z) = (z - a)^{-1}$ with complex a and some applications to certain non-hermitian matrices*, Approx. Theory Appl., 5 (1989), pp. 15–31.
- [18] NICHOLAS J. HIGHAM, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.
- [19] MARLIS HOCHBRUCK AND CHRISTIAN LUBICH, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
- [20] ARIEH ISELES, *How large is the exponential of a banded matrix?*, New Zealand J. Math., 29 (2000), pp. 177–192.
- [21] LEONID KNIZHNERMAN AND VALERIA SIMONCINI, *A new investigation of the extended Krylov subspace method for matrix function evaluations*, Numer. Linear Algebra Appl., 17 (2010), pp. 615–638.
- [22] NICOLA MASTRONARDI, MICHAEL KWOK-PO NG, AND EUGENE E. TYRTYSHNIKOV, *Decay in functions of multiband matrices*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2721–2737.
- [23] THE MATHWORKS, INC., *MATLAB 7, r2013b* ed., 2013.
- [24] MATRIX MARKET, *A visual repository of test data for use in comparative studies of algorithms for numerical linear algebra*, Mathematical and Computational Sciences Division, National Institute of Standards and Technology; available online at <http://math.nist.gov/Matrix-Market>.
- [25] GÉRARD MEURANT, *A review on the inverse of symmetric tridiagonal and block tridiagonal matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 707–728.
- [26] ROBERT F. RINEHART, *The equivalence of definitions of a matrix function*, Amer. Math. Monthly, 62 (1955), pp. 395–414.
- [27] YOUSEF SAAD, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
- [28] VALERIA SIMONCINI, *Variable accuracy of matrix-vector products in projection methods for eigencomputation*, SIAM J. Numerical Analysis, 43 (2005), pp. 1155–1174.
- [29] VALERIA SIMONCINI AND DANIEL B. SZYLD, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM J. Sci. Comput., 25 (2003), pp. 454–477.
- [30] PAVEL K. SUETIN, *Series of Faber polynomials*, Gordon and Breach Science Publishers, 1998. Translated from the 1984 Russian original by E. V. Pankratiev [E. V. Pankrat’ev].
- [31] HAO WANG, *The Krylov Subspace Methods for the Computation of Matrix Exponentials*, PhD thesis, Department of Mathematics, University of Kentucky, 2015.
- [32] HAO WANG AND QIANG YE, *Error Bounds for the Krylov Subspace Methods for Computations of Matrix Exponentials*, ArXiv e-prints, (2016). arXiv:1603.07358.
- [33] QIANG YE, *Error bounds for the Lanczos methods for approximating matrix exponentials*, SIAM J. Numer. Anal., 51 (2013), pp. 68–87.